Autistic Spectrum Disorder Screening Data for Toddlers – Data Wrangling and Preliminary Analysis

In first looking at the data set for my first Capstone project, Autistic Spectrum Disorder Screening Data for Toddlers, I wanted to clean up the column names. I replaced all spaces, ' ', with underscores, '_', then removed underscores that appeared at the end of a column name using a regular expression.

This was the only issue of this type with this data set. There were no problems with missing values, values of different types in the same series / column, outliers, etc. Essentially, this is a clean data set.

Since the data didn't involve extensive cleaning, I looked at some summary statistical information. I started with looking at classic Autism Spectrum Disorder (ASD) traits, Class/ASD_Traits, (Yes/No) by gender. As I expected from looking at past research, boys had a higher rate of Autism traits than girls. Boys also comprised more than twice as many subjects in this data.

The minimum age of subjects in this study was 12 months and the maximum was 36. I took a look at classic ASD traits (Yes/No) by age. The highest rate of a classification of yes appears to be at 23 months. I created an Age Range value grouping the ages into five ranges. The age range of 22 to 26 months has the highest rate of a positive response to classic ASD traits.

The Qchat-10-Score is the sum of the individual binary scores that are observed behavioral features associated with ASD. Basically the higher this number, the more likely the Class/ASD_Traits will have a value of Yes. The age group with the highest average Qchat-10-Score is 27-31 months. In looking at the average of this score by ethnicity, Pacific Islanders and Native Americans have the highest score.

I will look at these categories more closely as I get more in depth with the statistical analysis of this study.